Pergamon

Appl. Math. Lett. Vol. 10, No. 4, pp. 19–24, 1997

0893-9659/97 $17.00 + 0.00

PII: S0893-9659(97)00053-0

Monotone Linkage Clustering and Quasi-Concave Set Functions

Y. KEMPNER
School of Mathematical Sciences, Tel-Aviv University
Ramat-Aviv 69978, Israel

B. MIRKIN*
DIMACS, Rutgers University
P.O. Box 1179, Piscataway, NJ 08855-1179, U.S.A.
and
CEMI of Russian Academy of the Sciences, Moscow, Russia

I. MUCHNIK
RUTCOR, Rutgers University
New Brunswick, NJ 08903, U.S.A.

(Received December 1996; accepted January 1997)

Communicated by A. Dress

Abstract—Greedily seriating objects one by one is implicitly employed in many heuristic clustering procedures which can be described in terms of a linkage function measuring entity-to-set dissimilarities. A well-known clustering technique, single linkage clustering, can be considered as an example of the seriation procedures (based actually on the minimum spanning tree construction) leading to the global maximum of a corresponding "minimum split" set function. The purpose of this work is to extend this property to the wide class of so-called monotone linkages. It is shown that the minimum split functions of monotone linkages can be maximized greedily. Moreover, this class of set functions is proven to coincide with the class of so-called quasi-concave set functions.

Keywords—Clustering, Monotone linkage, Quasi-concavity, Greedy optimization.

1. INTRODUCTION

The subject of this paper originated in cluster analysis as a generalization of a well-known method called single linkage clustering [1]. Let $D = (d_{ij})$ be a symmetric $N \times N$ matrix of the dissimilarities d_{ij} between elements, i, j, of an N-element set I. For a nonempty subset $S \subset I$ and an element $i \in I - S$, let us define $l(i, S) := \min_{j \in S} d_{ij}$, the single linkage dissimilarity between i and S. A sequence $s = (i_1, i_2, \ldots, i_N)$ consisting of all elements of I will be referred to as a *series*, and the sets $S_k := \{i_1, \ldots, i_k\}$ consisting of the initial fragments of s, as its *starting sets* ($k = 1, 2, \ldots, N-1$). A series $s = (i_1, i_2, \ldots, i_N)$ is a *single linkage series* if, for every $k = 1, \ldots, N-1$, the element i_{k+1} is a minimizer of $l(i, S_k)$ with regard to $i \in I - S_k$. A starting set S_k of a single linkage series s can be referred to as a *single linkage cluster* if it is maximally separated from the other elements along that series, that is, if $l(i_{k+1}, S_k)$ is maximum over all

The authors are indebted to A. Dress for his numerous revising suggestions. B. Mirkin thanks the Office of Naval Research for its support under Grants Numbers N00014-93-1-0222 and N00014-96-1-0208 to Rutgers University. I. Muchnik gratefully acknowledges the support of DIMACS (Grant NSF-CCR-91-1999) and RUTCOR (Grant F49620-93-1-0041).
*Author to whom all correspondence should be addressed.

$k = 1, \ldots, N-1$. Basically, a single linkage series $s = (i_1, i_2, \ldots, i_N)$ defines a minimum spanning tree (MST) of the graph whose vertex set is I and whose edge weight function is $d = (d_{ij})$, according to the framework of the well-known Dijkstra-Prim algorithm for finding an MST; the edges of that MST connect, for $k = 1, \ldots, N-1$, the vertex i_{k+1} with just one of the vertices $j \in S_k$ with $d_{i_{k+1}j} = l(i_{k+1}, S_k)$. By cutting any MST at all its maximum weight edges (j, i_{k+1}), the set I is partitioned into classes that are inclusion-minimal single linkage clusters; moreover, a subset $S \subset I$ is a single linkage cluster if and only if it is a union of some of the maximum-weight-edge-cut classes. Also, a set function, $L(S) := \min_{i \in I-S} l(i, S)$, called the *minimum split function*, is maximized by every single linkage cluster [2].

The single linkage $l(i, S)$ satisfies a monotonicity property [3]: its value can only decrease when some elements (not coinciding with i) are added to S. All the clustering concepts above can be extended to an arbitrary monotone linkage function, $d(i, S)$, whether it is defined in terms of a dissimilarity matrix or not. This paper is aimed at proving that, for any monotone linkage function, d, its minimum split function, $M_d(S) := \min_{i \in I-S} d(i, S)$, has all its inclusion-minimal maximizers (over the set of nonempty proper subsets of I) among the "single linkage" clusters defined via the "single linkage" series. Moreover, it appears that the entire stock of the minimum split functions for the monotone linkages coincides with the set of quasi-concave set functions F defined by the condition

$$F(S_1 \cap S_2) \geq \min(F(S_1), F(S_2)),$$

for any overlapping $S_1, S_2 \subset I$.

This provides both a simple algorithm for maximizing quasi-concave set functions presented as minimum split functions for monotone linkage and a natural mechanism for generating quasi-concave set functions.

The remainder consists of the following. In Section 2, the monotone linkage functions are introduced and minimum split and maximum join concepts are analyzed; in particular, their relation with quasi-concavity is stated. In Section 3, it is shown that the minimal maximizers of a minimum split function are starting sets of the corresponding linkage series, while every nonminimal maximizer is just a union of some of the minimal ones. An example is given in Section 4. The results are discussed in Section 5.

2. MONOTONE LINKAGE AND QUASI-CONCAVE SET FUNCTION

In this paper, we use symbol $\mathcal{P}^-(I)$ to denote the set of all nonempty proper subsets of I so that $\emptyset, I \notin \mathcal{P}^-(I)$. Let, for every $S \in \mathcal{P}^-(I)$ and $i \in I - S$, a dissimilarity measure, $d(i, S)$, be given. Such a measure, referred to as a *linkage* between i and S, can be defined in terms of different data formats. For example, for a data table $X = (x_{ik})$ where x_{ik} is the value of a variable $k \in K$ for any entity $i \in I$, a linkage measure can be defined as

$$ml(i, S) := \sum_{k \in K} \min_{j \in S} |x_{ik} - x_{jk}|. \tag{1}$$

The *ml* linkage is an example of a "holistic" measure that cannot be reduced to a function of pairwise dissimilarities.

Let us refer to a linkage function $d(i, S)$, $S \in \mathcal{P}^-(I)$, $i \in I - S$, as a *monotone linkage* if $d(i, S) \geq d(i, T)$ whenever $S \subseteq T$ (for all $i \in I - T$). Both of the specific linkage functions considered, $l(i, S)$ and $ml(i, S)$, are monotone.

Given a linkage function d, a set function M_d can be defined on $\mathcal{P}^-(I)$ as follows:

$$M_d(S) := \min_{i \in I-S} d(i, S). \tag{2}$$

This set function was considered, among others, in [4]. Following the terminology introduced in [2], M_d can be referred to as the *minimum split function* for the linkage function d. The

minimum split function measures the minimum linkage between S as a whole and $I - S$ as the set of the "individual" entities. A set function $F : \mathcal{P}^-(I) \to R$ is called *quasi-concave* [4] if

$$F(S_1 \cap S_2) \geq \min(F(S_1), F(S_2)), \tag{3}$$

for any overlapping $S_1, S_2 \in \mathcal{P}^-(I)$.

ASSERTION 1. *The minimum split function of any monotone linkage is quasi-concave.*

PROOF. Let $F(S) := M_d(S) = \min_{i \in I-S} d(i, S)$ for some monotone linkage d, and let S_1, S_2 be overlapping elements of $\mathcal{P}^-(I)$. Assume $F(S_1 \cap S_2) = d(i, S_1 \cap S_2)$, $F(S_1) = d(j, S_1)$, and $F(S_2) = d(k, S_2)$. By the definition of F, i does not belong to at least one of S_1, S_2, say $i \notin S_1$. Then, $d(i, S_1) \geq F(S_1) = d(j, S_1)$ and $F(S_1 \cap S_2) = d(i, S_1 \cap S_2) \geq d(i, S_1)$ due to monotonicity of d, which proves that F is quasi-concave. ∎

Let us define now the *maximum join* linkage function d_F for any set function $F : \mathcal{P}^-(I) \to R$ as

$$d_F(i, S) := \max_{S \subseteq T \subseteq I-i} F(T), \tag{4}$$

for any $S \in \mathcal{P}^-(I)$ and $i \in I - S$.

ASSERTION 2. *The maximum join linkage d_F is monotone.*

PROOF. Obvious since any increase of S makes the set of maximized values in (4) smaller. ∎

Next, we show that in the setting defined by the conditions of quasi-concavity and monotonicity, there is a relationship between the functions d_F and M_d. For any quasi-concave set function $F : \mathcal{P}^-(I) \to R$, the minimum split function of its maximum join linkage coincides with F (Assertion 3). A much weaker property holds for the linkage functions: the maximum join linkage of the minimum split function of a monotone linkage d is not larger than d (Assertion 4).

ASSERTION 3. *For any quasi-concave set function $F : \mathcal{P}^-(I) \to R$, the minimum split function of its maximum join linkage coincides with F.*

PROOF. For an $S \in \mathcal{P}^-(I)$ and $i \in I - S$, let S_i be a maximizer of $F(T)$ over all T satisfying the condition $S \subseteq T \subseteq I - i$, so that $d_F(i, S) = F(S_i)$. The minimum split function for d_F, by definition, is equal to $M(S) := \min_{i \notin S} F(S_i)$. Thus, $M(S) \leq F(\cap_{i \notin S} S_i)$, due to quasi-concavity. But $\cap_{i \notin S} S_i = S$ since $S \subseteq S_i$ and $i \notin S_i$, for every $i \notin S$, which implies $M(S) \leq F(S)$. On the other hand, $F(S_i) \geq F(S)$ for any $i \notin S$ since S belongs to the set of feasible subsets in the definition of S_i as a maximizer of F; this implies that $M(S) \geq F(S)$, which proves the statement. ∎

Different linkage functions d and d' may produce coinciding minimum split functions, $M_d = M_{d'}$. The maximum join linkage is peculiar: it is the minimum in its class.

ASSERTION 4. *If a set function F is the minimum split function for a monotone linkage d, then $d_F(i, S) \leq d(i, S)$ for any $S \in \mathcal{P}^-(I)$ and $i \notin S$.*

PROOF. For an arbitrary $S \in \mathcal{P}^-(I)$ and any $i \in I - S$, assume $d_F(i, S) = F(T)$ for some T with $S \subseteq T \subseteq I - i$. By definition, $F(T) = \min_{j \in I-T} d(j, T) \leq d(i, T)$ since $i \in I - T$. However, $d(i, T) \leq d(i, S)$ since $S \subseteq T$ and d is monotone. Thus, $d_F(i, S) \leq d(i, S)$. ∎

There is also an algorithmic asymmetry between the concepts introduced: it is quite easy to construct the minimum split function M_d associated with a linkage d, while determining the maximum join linkage d_F for a set function F may be an exponentially hard problem: the former task involves the elements $i \in I - S$ to enumerate, while the latter requires maximizing a set function $F(T)$.

3. MAXIMIZING MINIMUM SPLIT QUASI-CONCAVE SET FUNCTION

In this section, we analyze the problem of maximizing quasi-concave set functions on $\mathcal{P}^-(I)$. Let us consider a quasi-concave set function F such that $F = M_d$ in (2) for a monotone linkage function d. Let us refer to a series $(i_1, \ldots, i_N)$ as a *d-series* if $d(i_{k+1}, S_k) = \min_{i \in I - S_k} d(i, S_k) = F(S_k)$ for any starting set $S_k = \{i_1, \ldots, i_k\}$, $k = 1, \ldots, N-1$. This definition describes a greedy procedure for constructing a d-series starting with $i_1 \in I$: having defined S_k, take any i minimizing $d(i, S_k)$ over all $i \in I - S_k$ as i_{k+1}, $k = 1, \ldots, N-1$. A subset $S \in \mathcal{P}^-(I)$ will be referred to as a *d-cluster* if there exists a d-series, $s = (i_1, \ldots, i_N)$, such that S is a maximizer of $F(S)$ over all starting sets S_k of s. Greedily found, d-clusters play an important part in maximizing the associated quasi-concave set function.

ASSERTION 5. *If, for a d-series $s = (i_1, i_2, \ldots, i_N)$, a subset $S \subset I$ contains i_1, and i_{k+1} is the first element in s not contained in S (for some $k = 1, \ldots, N-1$), then*

$$F(S_k) = d(i_{k+1}, S_k) \geq d(i_{k+1}, S) \geq F(S),$$

where $S_k = \{i_1, \ldots, i_k\}$. In particular, if S is an inclusion-minimal maximizer of F (with regard to $\mathcal{P}^-(I)$), then $S = S_k$, that is, S is a d-cluster.

PROOF. Indeed, $F(S_k) = d(i_{k+1}, S_k)$ by definition; $d(i_{k+1}, S) \geq d(i_{k+1}, S_k)$ by monotonicity; $d(i_{k+1}, S_k) \geq F(S_k)$ because $F(S_k) = \min_{i \in I - S_k} d(i, S_k)$ and $i_{k+1} \notin S_k$. ∎

ASSERTION 6. *If $S_1, S_2 \subset I$ are overlapping maximizers of a quasi-concave set function $F(S)$ over $\mathcal{P}^-(I)$, then $S_1 \cap S_2$ is also a maximizer of $F(S)$.*

PROOF. Obviously follows from (3). ∎

This means that the set of all maximizers of F in $\mathcal{P}^-(I)$ is a semilattice (with regard to set-theoretic inclusion and intersection). Assertion 6 implies also that the minimal maximizers of a quasi-concave set function over $\mathcal{P}^-(I)$ are not overlapping. Moreover, any nonminimal maximizer can be uniquely partitioned into a set of the minimal ones.

ASSERTION 7. *Each maximizer of a quasi-concave set function (on $\mathcal{P}^-(I)$) is a union of its inclusion-minimal maximizers.*

PROOF. Indeed, if S^* is a maximizer of $F = M_d$ over $\mathcal{P}^-(I)$, then, according to Assertion 5, for any $i \in S^*$, there exists a minimal maximizer included in S^* and containing i. ∎

It follows that we can find all minimal maximizers of a quasi-concave set function $F = M_d$ on $\mathcal{P}^-(I)$) for a monotone linkage d using the following three-step *extended greedy procedure* (EGP).

(A) For each $i \in I$, define some d-series p_i greedily starting from i as its first element.

(B) For each d-series $p_i = (i_1 := i, i_2, \ldots, i_N)$, let T_i denote its smallest starting set with $F(T_i) = \max_{1 \leq k \leq N-1} d(i_{k+1}, \{i_1, \ldots, i_k\})$.

(C) Among the noncoinciding minimal d-clusters T_i, $i \in I$, choose those maximizing F.

Performing EGP takes $O(N^3 g)$ time, where g is the average time required to calculate any single value $d(i, S)$ since, at Step (A), N series are constructed taking $O(N^2 g)$ time each.

ASSERTION 8. *The extended greedy procedure EGP finds all the minimal maximizers over $\mathcal{P}^-(I)$.*

PROOF. Assume that, for some $i \in I$, there exists a d-series q_i starting with i, whose minimal d-cluster Q_i does not belong to the set of clusters found with EGP. Then, $T_i \cap Q_i$ contains i and, thus, is a maximizer of F, strictly included in T_i, which contradicts the minimality of T_i. ∎

4. EXAMPLE

Let us consider the set I of rows of a 6×7 Boolean matrix X:

$$X := \begin{array}{c|ccccccc} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 2 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 3 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 4 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 5 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 6 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{array}.$$

The row-to-row Hamming distances (numbers of noncoinciding components) form the matrix

$$D := \begin{pmatrix} 0 & 4 & 3 & 7 & 4 & 3 \\ 4 & 0 & 5 & 3 & 2 & 5 \\ 3 & 5 & 0 & 4 & 5 & 4 \\ 7 & 3 & 4 & 0 & 3 & 4 \\ 4 & 2 & 5 & 3 & 0 & 3 \\ 3 & 5 & 4 & 4 & 3 & 0 \end{pmatrix}.$$

There are five minimal maximizers of the associated minimum split single linkage function L, as can be seen from the D-based MST, presented in Figure 1: $\{1\}$, $\{2,5\}$, $\{3\}$, $\{4\}$, $\{6\}$, obtained by cutting all the MST edges with maximum weight 3.

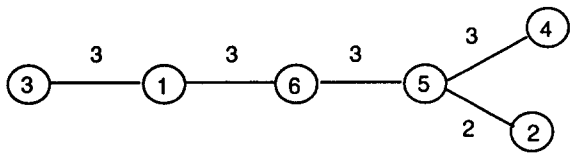

Figure 1. A minimum spanning tree for matrix D.

The situation is somewhat different for the minimum split of the "holistic" linkage ml defined in (1): its minimal maximizers (over $\mathcal{P}^-(I)$) are $\{1\}$, $\{3\}$, $\{4\}$, and $\{6\}$, while none of the elements 2 and 5 belongs to a maximizer of M_{ml}. Indeed, let us take a look at the following six ml-series starting from each of the elements of I: 1(3)3(3)2(0)5(1)4(0)6, 2(2)5(2)4(2)6(1)1(0)3, 3(3)1(3)2(0)5(1)4(0)6, 4(3)2(1)5(2)6(1)1(0)3, 5(2)2(2)(2)6(1)1(0)3, 6(3)1(2)3(2)2(0)4(0)5. The value $ml(i_{k+1}, S_k)$ is put in the parentheses between every starting set S_k seriated and i_{k+1} $(k = 1, \ldots, 5)$. It can be seen that the maximum value 3 separates each of the four singletons above while it never occurs in the series starting with 2 or 5.

5. CONCLUSION

The monotone linkage functions have been introduced, in the framework of clustering, by Mullat [3] who called them "monotone systems" and considered set functions $G(S) := \max_{i \in S} d(i, S)$ as greedily minimizable. In this paper, the concept of a minimum split function [2] is extended to the case of monotone linkage functions. We have proven that the inclusion-minimal maximizers of a minimum split function are monotone linkage clusters, all of which can be found with the extended greedy procedure EGP. This allows us to claim that the minimum split functions M_d for monotone linkages d present yet another class of greedily maximizable functions, though the greediness here is associated with the definition of M_d via d rather than with direct maximization of M_d considered usually (as, for instance, in [5]). We have proven also that this class coincides with the class of quasi-concave set functions.

Although the problem of maximizing quasi-concave set functions is exponentially hard when they are oracle-defined [6], it can be resolved with the extended greedy procedure when they are

defined in terms of a monotone linkage. Thus, the monotone linkage format may well serve as an easy-to-interpret and easy-to-maximize input for dealing with quasi-concave set functions.

On the other hand, the monotone linkage concept may be used as a framework for developing clustering techniques based on entity-to-set linkages rather than on conventional entity-to-entity dissimilarity measures. The "unclusterable," "noisy" entities frequently occurring in real-world data can be treated explicitly in this framework.

The constructions described only involve ordering information in both the domain and range of set/linkage functions, and also, they rely on the fact that every subset is uniquely decomposable into its elements. Therefore, they can be extended to distributive lattice structures considering the set of irreducible elements as I (see, for instance, [7] where relations between monotonicity and quasi-concavity on distributive lattices and semilattices have been studied).

REFERENCES

1. P.H.A. Sneath and R.R. Sokal, *Numerical Taxonomy*, W.H. Freeman, San Francisco, (1973).
2. M. Delattre and P. Hansen, Bicriterion cluster analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **4**, 277–291 (1980).
3. J. Mullat, Extremal subsystems of monotone systems: I, II, *Automation and Remote Control* **37**, 758–766; 1286–1294 (1976).
4. Yu. Zaks (Kempner) and I. Muchnik, Incomplete classifications of a finite set of objects using monotone systems, *Automation and Remote Control* **50**, 553–560 (1989).
5. A.W.M. Dress and W. Terhalle, Well-layered maps—A class of greedily optimizable set functions, *Appl. Math. Lett.* **8** (5), 77–80 (1995).
6. V. Levit, Oracle-defined quasi-concave set functions are exponentially hard to maximize, Personal communication (1995).
7. L. Libkin, I. Muchnik and L. Shvartser, Quasi-linear monotone systems, *Automation and Remote Control* **50**, 1249–1259 (1989).